CONSENSUS BASED MEASUREMENT

P. J. Legree and J. Psotka
U.S. Army Research Institute for the Behavioral and Social Sciences
Arlington, VA 22202

Abstract

Developing and scoring situational judgment tests have usually required much expert opinion. A more powerful, broader, and still cost-efficient procedure for creating standards even in ill-defined domains, termed *Consensus Based Measurement* (CBM), allows examinee responses to be evaluated as deviations from consensus understandings implied by the response distributions of examinee samples. Evaluative data show substantial convergence between expert and examinee based standards and scores, and indicate CBM may be used to score SJTs even when expert judgments are not available to develop scoring rubrics.

1. Background

The Army uses situational judgment technologies and materials to improve supervisory, leadership, and interpersonal knowledge, skills, and values that affect Soldier performance, and it is likely that the importance of these human characteristics will increase as units continue to become more autonomous, flexible, and powerful (cf., Hedlund et al., 2003). Closely related assessment center technologies have been utilized for industrial and scientific purposes to develop models of performance and evaluate theories of cognition (Mayer, Caruso & Salovey, 1999; McDaniel et al., 2000). Therefore, technologies supporting situational judgment tests' development have both practical importance for Army operations and scientific importance for psychologists.

Situational judgment is required in many practical situations that individuals encounter in their personal life and in job-related settings, and superior performance in these situations often requires knowledge reflecting a wide range of experiences. Situational judgment tests (SJTs) have been constructed to describe these situations. These scales require examinees to endorse either actions or interpretations that might be associated with the simulated event. SJTs have been described as low fidelity simulations because ambiguity is necessarily associated with the situations, actions and interpretations. Assessing performance on these scales requires the development of scoring rubrics that are sensitive to this ambiguity.

To ensure relevance to the performance domain, the development of SJTs has traditionally required much expert judgment to: (a) identify and describe situations, (b) specify relevant interpretations and responses, and (c) develop scoring rubrics to assess performance on the instruments. These scales often assess abilities in soft domains, such as interpersonal and supervisory skills, to support personnel selection and development. This approach has been problematic because while substantial numbers of experts are required for scale development, sometimes experts have been difficult to identify, may have competing time requirements, or may provide inconsistent information. In addition, some domains lack certified experts, and the specification of knowledge for emerging domains may be incomplete and impossible through expert opinion.

2. Consensus Based Measurement

A simpler, cost-efficient procedure, termed *Consensus Based Measurement* (CBM), can, and more broadly should be used even when experts are available. This approach leverages models of human performance by postulating that errors in opinions are random and not systematic over individuals (cf. Legree, Psotka, Tremble & Bourne, in press; Legree 1995). CBM is particularly well suited for those cases in which expertise is rare or difficult to identify and for emerging domains for which understandings may not have been well-specified.

Our conceptualizations regarding CBM evolved from expectations about how item response distributions might change as a function of the expertise of respondent samples. Knowledge is customarily viewed as growing over levels of expertise within any specific domain. Therefore, if a sample of apprentices were tracked over time, and repeatedly surveyed with standard knowledge items as novices, journeymen, and experts, the response distributions in Figure 1 might describe their growth in expertise. The distributions in Figure 1 illustrate both individual differences and increasing knowledge.

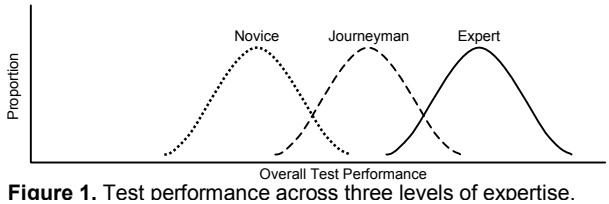

Figure 1. Test performance across three levels of expertise.

However, suppose supervisors were surveyed with items that required endorsement of statements using a Likert scale. For example, supervisors might be requested to rate the importance of maintaining morale to support team performance. For this type of item, the response distributions associated with increased levels of expertise (i.e., those supervisors who are more knowledgeable) might

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
00 DEC 2004	**N/A**	**-**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Consensus Based Measurement	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **U.S. Army Research Institute for the Behavioral and Social Sciences Arlington, VA 22202**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release, distribution unlimited

13. SUPPLEMENTARY NOTES
See also ADM001736, Proceedings for the Army Science Conference (24th) Held on 29 November - 2 December 2005 in Orlando, Florida.

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**UU**	**2**	

vary in both central tendency and in variance. A change in central tendency, which is illustrated in Figure 2a, would occur as individuals learn that maintaining morale may carry indirect implications for performance. A reduction in variance might occur as respondent understandings concerning morale become more refined, allowing recognition that while morale carries implications for team performance, these implications may be limited. Figure 2b illustrates a reduction in variance of response distributions associated with increased accuracy.

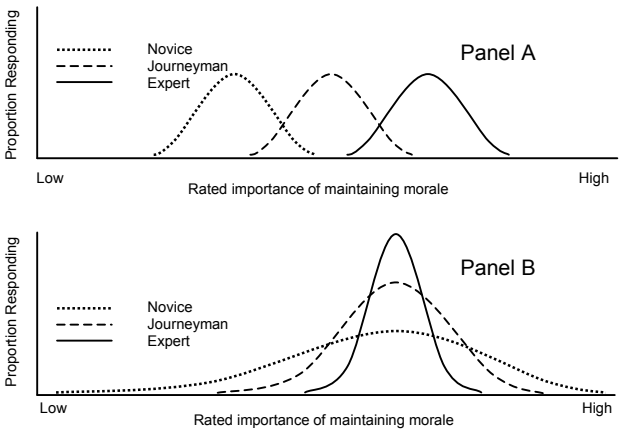

Figure 2. Likert item responses across levels of expertise.

Both these trends may have general relevance to understanding the growth and refinement of knowledge. By definition, naïve individuals have poorly formed conceptual structures for understanding relationships or events, and their responses may not be sensible, sometimes indicating ignorance of even basic relationships and sometimes overstating their importance. However, with increasing degrees of sophistication, individuals become increasingly aware and accurate in their understandings of relationships and events. To the extent poor performance on a knowledge test can be viewed as reflecting error, non-expert responses will be more variable than those of experts, as well as possibly having a different central tendency.

These conceptualizations suggests that by phrasing items in the form of Likert items, mean expert ratings might be approximated by mean journeymen ratings. Substantial convergence (Figure 2b) would occur when the main difference across levels of expertise corresponds to differences in variance as opposed to central tendency, and the assessment of this possibility, if endorsed, would allow the development of scales for domains without the necessity of expert opinion data.

3. Results & Conclusions

To evaluate these conceptualizations, four datasets were identified that support the assessment of examinee responses using traditional expert-based scoring as well as CBM. The level of convergence between both scoring rubrics and scores was computed for each dataset as the correlation between sets of values.

Table 1 summarizes the level of convergence between both the scoring rubrics and the resultant scores for those datasets. These results show substantial convergence between situational judgment tests scored using expert and examinee based scoring standards computed without reference to criterion data for which substantial expert and examinee data are available. The analyses indicate that CBM may be used to develop and score situational judgment tests when expert responses are not available or of limited quality. This technology is ideal for identifying knowledge in emerging domains that have not been well-specified, are dynamic, or may lack any experts.

Data that provide evidence in support of the additional hypothesis that CBM in many circumstances is superior to expert – generated rubrics is advanced in Legree et al. (In Press).

Table 1. Summary results from four datasets supporting expert and consensus based scoring.

Scale / Source	Scoring Key convergence	Score convergence
Project A SJT (Legree, 1995)	.74	.88
MSCEIT (Mayer Caruso & Salovey, 1999)	.90	.98
TKML (Legree, Psotka, Tremble & Bourne, in press)	.96	1.00
NCO21Supervisory SJT (Heffner & Porr, 2003)	.89	.95

4. References

Legree, P. J. (1995). Evidence for an oblique social intelligence factor. *Intelligence, 21,* 247-266.

Legree, P. J., Martin, D., & Psotka, J. (2000). Measuring cognitive aptitude using unobtrusive knowledge tests: A new survey technology. *Intelligence*, *28*, 291–308.

Legree, P. J., Psotka, J., Tremble, T. & Bourne, D. (in press). Using Consensus Based Measurement to Assess Emotional Intelligence. In R. Schulze & R. Roberts (Eds.) *International Handbook of Emotional Intelligence.*

Hedlund, J., Forsythe, G. B., Horvath, J. A., Williams, W. M., Snook, S., & Sternberg, R. J. (2003). Identifying and assessing tacit knowledge: Understanding the practical intelligence of military leaders. *Leadership Quarterly, 14,* 117-140.

Heffner, T. S. & Porr W. B. (2000, August). *Scoring Situational Judgment Tests: A Comparison of Multiple Standards Using Scenario Response Alternatives.* Paper presented at the Annual Conference of the APA, Washington, DC.

Mayer, J. D., Caruso, D. R., & Salovey, P. (1999). Emotional intelligence meets traditional standards for an intelligence. *Intelligence, 27,* 267-298.

McDaniel, M. A., Morgeson, F. P., Finnegan, E. B., Campion, M. A., & Braverman, E. P. (2001). Use of situational judgement tests to predict job performance: A clarification of the literature. *Journal of Applied Psychology, 86*, 730 – 740.